IJCRT.ORG ISSN : 2320-2882
INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

DETECTING PHISHING ATTACK USING NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING

Sangeetha M[1]. Giriswarna R[2]. Harini P[3]

[1]Associate Professor, Department of CSE, Panimalar Engineering College, Chennai.India [2] U.G Scholar, Department of CSE, Panimalar Engineering College, Chennai.India [3]U.G Scholar, Department of CSE, Panimalar Engineering College, Chennai.India

Abstract

Malicious websites largely promote the expansion of Internet criminal activities and constrain the event of Web services. As a result, there has been strong motivation to develop systemic solution to stopping the user from visiting such Websites. We propose a learning based approach to classifying websites into 3 classes: Benign, Spam and Malicious. Our mechanism only analyses the Uniform Resource Locator (URL) itself without accessing the content of websites. Thus, it eliminates the run-time latency then likelihood of exposing users to the browser based vulnerabilities. By employing learning algorithms, our scheme achieves better performance on generality and coverage compared with blacklisting service.

Keywords - Malicious, Vulnerabilities, Spam, Websites, Malware

I INTRODUCTION

Phishing has become one among the most deadly attacks. There are various approaches to threatening phishing attacks such as Phish net, lexical-based on-line learning, and a proactive phishing identification approach. In addition, traditional ways to discourage phishing email square measure through the use of vendor-based solutions like acanthopterygian Email Security entry and Symantec electronic communication entry, but each systems need email traffic redirection to every security appliance. Though trafficker solutions could determine phishing emails, they are doing not stop Associate in Nursing user from clicking on a malicious link inside a flagged email which will lead to compromising a computer system. To discourage such considerations, ones propose a range of Intrusion Detection Systems (IDS) and Intrusion Hindrance Systems (IPS) approaches to distinguishing and deterring phishing emails, however they lack feasibility or can't be used once e-mail communication becomes encrypted, that is usually done these days.

II LITERATURE SURVEY

Ram B. Basnet, Andrew H. Sung, in their project they used Naıve Bayes but they faced drawbacks because they don't use DNS entries and geo-locations of page's host and name servers 2014[1]. Wen Zhang, Yu-Xin Ding, Yan Tang, Bin Zhao, they used CV algorithm, PV algorithm, online learning algorithms but they faced the drawbacks such as burden of features is decided by the difference of feature frequency in malicious and benign samples, Experimental results show that this method can improve the performance of online learning algorithms,2011[2]. Ahmed, Abdullah, in their project they used content-based, heuristic-based and blacklist-based approaches but accuracy should be Improved,2016[3]. Birhanu Eshete, Adolfo Villafiorita, and Komminist Weldemariam, they used Confidence Weighted Majority Vote Classification, but one limitation of BINSPECT is lack of research of obfuscated JavaScript and emulation of the browser with plugins,2012[4]. J.Shad and S.Sharma, they used Heuristic,Blacklist,fuzzy rule based, image processing,CANTINA based approach but there is no classifiers used ,so less prediction,2018[5]. S.Marchal, J. Francois, R. State, and T. Engel, they used Naive bayes,So, Less accuracy, 2015[6]. Khonji,Iraqi,Andy Jones, in their project they used Signature technique,State-change technique (rule-based technique) some further work is completed to reinforce its Performance,2013[7]. M. Karabatak and T.Mustafa, they used Decision tree, Gradient Boosting, Generalised Addictive Model but we need more accuracy, 2018[8]. Wang Tao, Yu Shun Zheng, Xie Bailin, they used C4.5 algorithm,Svm,naïve bayes, Decision tree. Here the only drawback is they abandon the features relative with the analysis of web content, 2010[9]. Ping

Yi,Yuxiang Guan,Futai Zou,Yao Yao, Wei Wang, and Ting Zhu , they used Decision Tree but provides less performance, 2018[10].

III EXISTING SYSTEM

A poorly structured NN model may cause the model to under fit the training dataset. On the opposite hand, exaggeration in restructuring the system to suit every single item within the training dataset may cause the system to be over fitted. One possible solution to avoid the Over fitting problem is by restructuring the NN model in terms of tuning some parameters, adding new neurons to the hidden layer or sometimes adding a brand new layer to the network. For example, the model designer may set the appropriate error rate to a worth that's unreachable which causes the model to stay in local minima or sometimes the model designer may set the suitable error rate to a price which will further be improved

IV PROPOSED SYSTEM

Analysing lexical features enables us to capture the property for classification purposes. We first distinguish the 2 parts of a URL: the host name and also the path, from which we extract bag-of-words (strings delimited by '/', '?', '.', '=', '-' and '').We find that phishing website prefers to possess longer URL, more levels (delimited by dot), more tokens in domain and path, longer token. Besides, phishing and malware websites could pretend to be a benign one by containing popular brand names as tokens aside from those in second-level domain. Considering phishing websites and malware websites may use IP address directly so on cover the suspicious URL, which is extremely rare in benign case. Also, phishing URLs are found to contain several suggestive word tokens (confirm, account, banking, secure, ebayisapi, webscr, login, sign in).

V SYSTEM ARCHITECTURE

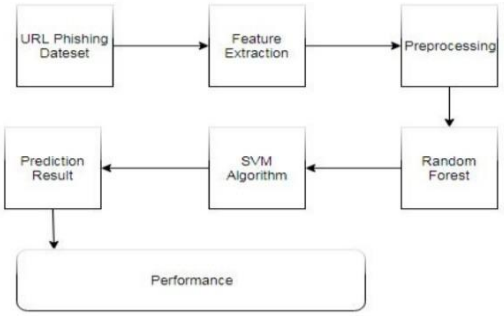

Fig 1 : Architecture to find Phishing website.

Fig 1 represents the architecture to find Phishing website. Initially we want to collect some dates in data set. In feature extraction we analyze the features of the input URL. Based on the features in preprocessing we will find whether the given URL is phishing website or non-phishing website by random forest and SVM algorithm.

VI FEATURES OF PHISHING WEBSITES

One of the challenges faced by our research was the unavailability of reliable training datasets. In fact, this challenge faces any researcher in the field. However, although plenty of articles about predicting phishing websites using data mining techniques have been disseminated these days, no reliable training dataset has been published publicly, maybe because there is no agreement in literature on the definitive features that characterize phishing websites, hence it is difficult to shape a dataset that covers all possible features. In this article, we shed light on the important features that have proved to be sound and effective in predicting phishing websites. In addition, we proposed some new features, experimentally assign new rules to some well-known features and update some other features.

Address Bar based Features

A. Features using IP Address

If an IP address is used as an alternative of the domain name in the URL, such as "http://125.98.3.123/fake.html", users can be sure that someone is trying to steal their personal information. Sometimes, the IP address is even transformed into hexadecimal code as shown in the following link

"http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html"

Rule: If The Domain Part has an IP Address then it is a Phishing website otherwise it is a Legitimate website

B. Long URL to Hide the Suspicious Part

Rule: If URL length < 54 character then it is a Legitimate website, else if URL length ≥ 54 and ≤ 75 then it is a Suspicious website, otherwise it is a Phishing website

C. Tiny URL to Hide the Suspicious Part

Rule: If TinyURL then it is a Phishing website, Otherwise it is a Legitimate website

D. URL having "@" Symbol

Using "@" symbol in the URL leads the browser to ignore everything preceding the "@" symbol and the real address often follows the "@" symbol.

Rule: If Url Having @ Symbol then it is a Phishing website, Otherwise it is a Legitimate website

E. URL having "//" Symbol

The existence of "//" within the URL path means that the user will be redirected to another website. An example of such URL's is: "http://www.legitimate.com//http://www.phishing.com". We examine the location where the "//" appears. We find that if the URL starts with "HTTP", that means the "//" should appear in the sixth position. However, if the URL employs "HTTPS" then the "//" should appear in seventh position.

Rule: If the Position of the Last Occurrence of "//" in the URL > 7 then it is a Phishing website, Otherwise it is a Legitimate website

F. URL having "-" Symbol

Rule: If Domain Name Part Includes (−) Symbol then it is a Phishing website, Otherwise it is a Legitimate website

G. Features based on Domain registration period

Based on the fact that a phishing website lives for a short period of time, we believe that trustworthy domains are regularly paid for several years in advance. In our dataset, we find that the longest fraudulent domains have been used for one year only.

Rule: If the Domains Expires on ≤ 1 years then it is a Phishing website, Otherwise it is a Legitimate website

VII BASED ON POP-UP WINDOW

It is unusual to find a legitimate website asking users to submit their personal information through a pop-up window. On the other hand, this feature has been used in some legitimate websites and its main goal is to warn users about fraudulent activities or broadcast a welcome announcement, though no personal information was asked to be filled in through these pop-up windows.

Rule: If Popup Window Contains Text Fields then it is a Phishing website, Otherwise it is a Legitimate website

VIII METHODS TO CLASSIFY WEBSITES

To ensure that our approach works well irrespective of the underlying classifier chosen for the task, we performed the experiments using two different classifiers: Random Forest and Support vector machine, as these are some of the most commonly used classifiers for the task of text-data classification. Scikit-learn implementation of these classifiers with their default parameter settings are used for our experiments. The tf-idf feature is used to represent each URL in the Database.

IX RANDOM FOREST AND SVM

Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction which shows in fig 2. SVM is used to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future that shows in fig 3.

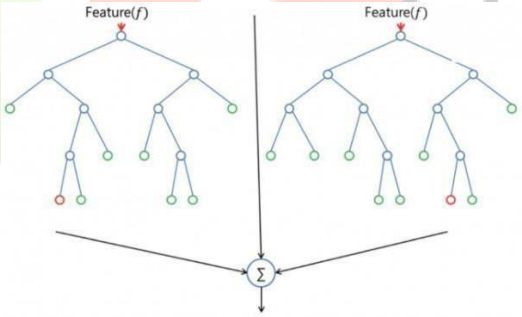

Fig 2 : Random Forest

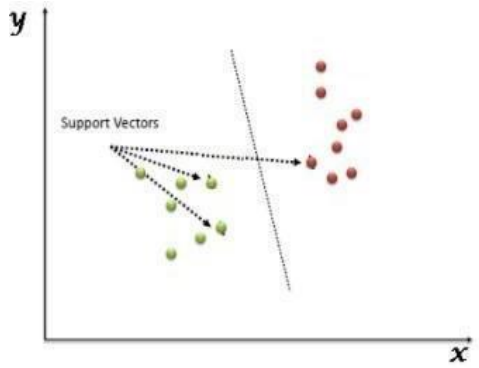

Fig 3 : Support Vector Machine

X CONCLUSION

In our project, all of URL in the dataset are labelled, which are done through Natural Language Processing especially during Feature Extraction. By using Machine Learning Algorithm Such as Random Forest and Support Vector Machine it takes less amount of time for execution and it gives more accuracy. Atlast, we can find whether the given website is a Phishing website or Legitimate Website.

XI REFERENCES

[1]. Ram B.Basnet , Andrew H.Sung , " Learning to detect Phishing URLs", 2014.
[2]."Malicious web Page detection based on on-line learning algorithm", Wen Zhang, Yu-Xin Ding, Yan Tang, Bin Zhao , 2011.
[3]. A. A. Ahmed and N. A. Abdullah, "Real time detection of phishing websites," 7th IEEE Annu. Inf. Technol. Electron. Mob. Commun. Conf. IEEE IEMCON 2016, 2016.
[4].Holistic Analysis and Detection of Malicious Web pages - Birhanu Eshete, Adolfo Villafiorita , and Komminist Weldemariam , 2012.
[5]. J. Shad and S. Sharma, "A Novel Machine Learning Approach to Detect Phishing Websites Jaypee Institute of Information Technology," pp. 425–430, 2018.
[6]. S. Marchal, J. Francois, R. State, and T. Engel, "Phish Score: Hacking phishers' minds," Proc. 10th Int. Conf. Netw. Serv. Manag. CNSM 2014, pp. 46–54, 2015.
[7]. Phishing Detection: A Literature Survey - Khonji,Iraqi,Andy Jones , 2013 .
[8]. Performance comparison of classifiers on reduced phishing website dataset - M. Karabatak and T. Mustafa , 2018.
[9]. A Novel Framework for learning to Detect Malicious web Pages - Wang Tao, Yu Shun Zheng, Xie Bailin , 2010
[10]. Web Phishing Detection Using a Deep Learning Framework Ping Yi,YuxiangGuan,Futai Zou,Yao Yao, Wei Wang, and Ting Zhu , 2018.
[11]."Detecting Phishing Websites via Aggregation Analysis of Page Layouts", JianMao,JingdongBian,WenqianTian,ShishiZhu, TaoWei, AiliLi, ZhenkaiLiang, 2018.
[12].Phishing website detection based on effective Machine Learning Approach",Gururaj Harinahalli Lokesh, Goutham, 2019.
[13].S. Afroz and R. Greenstadt, "Phish zoo: An automated web phishing detection approach based on profiling and fuzzy matching," Technical Report DU-CS-09-03, Drexel University, Tech. Rep., 2009.
[14].K.-T. Chen, J.-Y. Chen, C.-R. Huang, and C.-S. Chen, "Fighting phishing with discriminative keypoint features," IEEE Internet Computing, vol. 13, no. 3, pp. 56–63, 2009.
[15]. I. Fette, N. Sadeh, and A. Tomasic, "Learning to detect phishing emails," in WWW2007: Proceedings of the 16th International World Wide Web Conference, 2007.